A Note on the Article 'Inference for multivariate normal mixtures' by J. Chen and X. Tan

Grigory Alexandrovich[1]

Department of Mathematics and Computer Science, Marburg University, Germany

The current note discusses the consistency proof for the penalized maximum likelihood estimator of a Gaussian mixture from the paper 'Inference for multivariate normal mixtures' by J. Chen and X. Tan. A soft spot in that proof is identified and a rigorous alternative proof based on a uniform law of iterated logarithm is given.

Keywords: *Gaussian Mixtures, penalized maximum likelihood estimation, Law of Iterated Logarithm for VC classes, strong consistency*

1 Introduction

Gaussian mixture models provide a flexible tool for data modeling, clustering and classification. We consider the problem of estimating the parameters of a multivariate Gaussian mixture with K components by maximizing the likelihood function. This approach has a theoretical drawback: the likelihood function is unbounded, and the interesting maxima are local maxima in the interior of the parameter space. Consider an estimator with $\hat{\mu}_1 = Y_1$, $|\hat{\Sigma}_1| = \varepsilon$, $\hat{\mu}_2 \in \mathbb{R}^d$ arbitrary, $\hat{\Sigma}_2 = I$, $\hat{p}_1 = 1/2$. Then the likelihood function tends to infinity as $\varepsilon \to 0$ and hence the MLE is not consistent.

Two basic strategies for overcoming the unboundedness were studied in the literature: restricted optimization and penalization of the likelihood. In the first case a lower bound on the variances or their ratios is imposed, see e.g. Hathaway [6]. In the second case a term which penalizes small variances or ratios of variances is added to the log likelihood, see e.g. Ciuperca et al. [5], Tanaka [8], Chen et al. [4], Chen and Tan [3]. The second approach has some advantages over the first one - there is no tuning constant to choose and the penalty function actually disappears with increasing sample size.

In the current paper I discuss the consistency proof of the penalized MLE from Chen and Tan [3]. Among the above papers on consistency of the penalized MLE it is the most interesting one in the context of Gaussian mixtures, since it treats the multivariate case. Adjusting the penalty magnitude is an important issue and requires an assessment

[1]Address for correspondence: G. Alexandrovich, Department of Mathematics and Computer Science, Marburg University, Hans-Meerwein-Strasse 6. D-35032 Marburg, Germany, Email: alexandrovich@mathematik.uni-marburg.de, Fon: +49 6421 28 25462

of the number of observations with a high likelihood contribution. Such an assessment is given in Lemma 2 in Chen and Tan [3]. However its proof seems to contain a soft spot and I was not able to fix it. In Section 2 I elaborate on the soft spot in detail. In Section 3 I give an alternative proof of a similar statement based on a uniform law of iterated logarithm. This allows to make Chen and Tan's nice consistency proof fully rigorous.

2 Outline of Chen and Tan's consistency proof

In the following let L_n be the log-likelihood, φ the normal density, Θ the set of K-component mixture parameters $(\mu_j, \Sigma_j, p_j,\ 1 \leq j \leq K)$, where $\mu \in \mathbb{R}^d,\ |\Sigma| \in \mathcal{P}_d,\ p \in [0,1],\ \sum p_j = 1$ and $\mathcal{P}_d$ is the set of $d \times d$ symmetric positive definite matrices. Two parameters are considered as equivalent if they induce the same distribution. θ_0 denotes a true parameter. The proof has roughly the following scheme:

1 Divide the parameter space Θ in $K+1$ disjoint subsets $\Gamma_1, \ldots, \Gamma_{K+1}$ where each subset is characterized by the number of components whose covariances are bounded away from zero. The subset where all covariances are bounded away from zero, Γ_{K+1}, is regular and contains the true parameter θ_0 so the classical MLE theory as in Wald [10] or Kiefer and Wolfowitz [7] can be applied.

2 Show that asymptotically the penalized MLE $\hat{\theta}_{n,pMLE}$ a.s. does not lie in any subset except the regular one, that is

$$\sup_{\theta \in \Gamma_i} L_n(\theta) + p_n(\theta) - L_n(\theta_0) - p_n(\theta_0) \to -\infty, \quad i \in \{1, \ldots, K\},$$

where $p_n : \Theta \to \mathbb{R}$ is a penalty function.

The second step is quite involved and will be outlined more precisely. The penalty function p_n fulfils several conditions, see Chen and Tan [3]. Recall the key condition C3: $\tilde{p}_n(\Sigma) \leq 4(\log n)^2 \log|\Sigma|$ for $|\Sigma| < cn^{-2d}$, where $p_n(\theta) = \sum_{j=1}^K \tilde{p}_n(\Sigma_j)$ and c some positive constant. This condition is imposed in order to rule out the damaging effect of components with degenerate covariance matrices. It will turn out, that $4(\log n)^2$ is actually not sufficient.

A key element of the proof is a uniform assessment of the number of observations, with a high contribution to the likelihood. These are observations, that are located in certain critical regions. It turns out that an appropriate choice for such critical regions are sets

$$\tilde{A}(\mu, \Sigma) := \{y \in \mathbb{R}^d\ :\ (y-\mu)^\intercal \Sigma^{-1} (y-\mu) \leq (\log|\Sigma|)^2\},$$

where μ and Σ correspond to a degenerate component of the point at which the likelihood is evaluated.

The contribution of observations inside such a set will be ruled out by the penalty function and the one outside can be shown to be small enough. Precisely, following bounds are used

$$\varphi(y,\mu,\Sigma) \le \begin{cases} |\Sigma|^{-\frac{1}{2}} & y \in \tilde{A}(\mu,\Sigma), \\ \exp(-\frac{1}{4}(y-\mu)^\mathsf{T}\Sigma^{-1}(y-\mu)) & \text{otherwise.} \end{cases}$$

A statement of the form

$$H_n(\mu,\Sigma) := \sum_{i=1}^{n} \mathbf{1}_{Y_i \in \tilde{A}(\mu,\Sigma)} \le a(n) + b(n,|\Sigma|), \tag{1}$$

for all $\mu \in \mathbb{R}^d,\ \Sigma \in \mathcal{P}_d$ almost sure, is needed, where $a(n) = o(n)$ and $b(n,s) = O(n)$ for each s and $b(n,s)\log s^{-1/2} \to 0$ as $s \to 0$. An important detail here is, that the *almost sure* statement has to hold simultaneously for all tuples (μ,Σ) and not solely for each one. Given any statement with these properties one can prove the consistency of the penalized MLE, if the penalty function fulfils a modified condition C3: $\tilde{p}_n(\Sigma) \le a(n)\log|\Sigma|$ for $|\Sigma| \le cn^{-2d}$.

Chen and Tan [3] claimed essentially the following bound (Lemma 2)

$$H_n(\mu,\Sigma) \le 4(\log n)^2 + 8nM|\Sigma|^{1/2d}\log|\Sigma|, \tag{2}$$

for all (μ,Σ) with $|\Sigma| < \exp(-4d)$ a.s., where M is an upper bound of the true mixture density. The proof uses an ascription to the univariate case, which was proved in Chen et al. [4] by applying a Bernstein inequality and Borel-Cantelli Lemma. We omit further details of this involved proof and refer to the source. Instead, we pay our attention on the ascription, which actually does not work. The argument behind the ascription is the following

$$\begin{aligned} \{y \in \mathbb{R}^d \ :\ & (y-\mu)^\mathsf{T}\Sigma^{-1}(y-\mu) \le (\log|\Sigma|)^2\} \\ &= \{y \in \mathbb{R}^d \ :\ \sum \lambda_j^{-1}|a_j^\mathsf{T}(y-\mu)|^2 \le (\log|\Sigma|)^2\} \\ &\subseteq \{y \in \mathbb{R}^d \ :\ |a_j^\mathsf{T}(y-\mu)| \le -\sqrt{\lambda_j}\log|\Sigma|,\ j=1,\dots,d\} \\ &\subseteq \{y \in \mathbb{R}^d \ :\ |a_j^\mathsf{T}(y-\mu)| \le -\sqrt{\lambda_1}\log|\Sigma|\}, \end{aligned}$$

where $a_1,\dots,a_d$ and $\lambda_1 \le \dots \le \lambda_d$ are unit length eigenvectors and the corresponding eigenvalues of Σ respectively.

Further one argues that for every bounded set $B \subset \mathbb{R}^d$, there exists a finite subset of the unit d-sphere $Q \subset S^{d-1}$, such that for every $a \in S^{d-1}$ there exists a $b \in Q$ with the following property

$$\{y \in B : |a^\mathsf{T}(y-\mu)| \le -\sqrt{\lambda_1}\log|\Sigma|\} \subseteq \{y \in B : |b^\mathsf{T}(y-\mu)| \le -\sqrt{2\lambda_1}\log|\Sigma|\} \tag{3}$$

and concludes

$$H_n(\mu,\Sigma) \le \max_{b\in Q} \sum_{i=1}^{n} \mathbf{1}_{\{|b^\mathsf{T}(Y_i-\mu)| \le -\sqrt{2\lambda_1}\log|\Sigma|\}},$$

for every $\Sigma \in \mathcal{P}_d$. Hence the problem is reduced to univariate normal samples $b^\mathsf{T} Y_1, \dots b^\mathsf{T} Y_n$ for finitely many $b \in S^{d-1}$. But the argument seems to be not fully rigorous, since the inclusion (3) holds only given a fixed, bounded set B but not on the whole $\mathbb{R}^d$. I found no easy way to correct the ascription to the univariate case. However, there is an alternative, more easy approach.

3 Approach based on the uniform law of iterated logarithm

For the next statements the term of the Vapnik-Chervonenkis dimension of a class of sets is needed. This combinatorial concept serves for characterization of the complexity of a class of sets.

Definition 1. Let $\mathcal{X}$ be a complete separable metric space, $\mathcal{C} \subset 2^{\mathcal{X}}$ a family of subsets, $D \subset \mathcal{X}$ any finite subset. The *shatter coefficient* of $\mathcal{C}$ with respect to D is defined by

$$S(D : \mathcal{C}) := |\{C \cap D \ : \ C \in \mathcal{C}\}|. \tag{4}$$

The *VC dimension* of $\mathcal{C}$ $dim(\mathcal{C})$ is the largest integer $k \in \mathbb{N}$ such that $S(D : \mathcal{C}) = 2^k$ for some k-element subset D of $\mathcal{X}$. If for every k there exists a finite k-element subset $D \subset \mathcal{X}$ such that $S(D : \mathcal{C}) = 2^k$, then $dim(\mathcal{C}) = \infty$.

A class $\mathcal{C}$ with a finite VC dimension is called a *VC class*.

A class $\mathcal{F}$ of real valued functions $\mathcal{X} \to \mathbb{R}$ is called a *VC-graph class* if the collection of all sub-graphs of the functions in $\mathcal{F}$ forms a VC class of sets in $\mathcal{X} \times \mathbb{R}$.

VC classes have some comfortable properties, like being Glivenko-Cantelli or even Donsker classes, see e.g. van der Vaart and Wellner [9].

If $\mathcal{C}$ is a VC class, then the class $\mathcal{F} := \{1_C \ : \ C \in \mathcal{C}\}$ of indicator functions is a VC-graph class satisfying conditions of Theorem 2.13 from Alexander [2] and the next statement follows.

Theorem 1. *Let $\mathcal{C} \subset \mathcal{B}^d$ be a VC class of sets, $(Y_n)_{n\in\mathbb{N}}$ a d-dimensional i.i.d. process. Then a.s.*

$$\limsup_{n\to\infty} \sup_{C\in\mathcal{C}} \frac{|\sum_{i=1}^n 1_C(Y_i) - nP_{Y_1}(C)|}{\sqrt{2n \log\log n}} = \sup_{C\in\mathcal{C}} (P_{Y_1}(C)(1 - P_{Y_1}(C)))^{1/2}. \tag{5}$$

Hence follows

Corollary 2. *Let $(Y_n)_{n\in\mathbb{N}}$ be a d-dimensional i.i.d. process and*

$$\mathcal{E}_d := \left\{ \{y \in \mathbb{R}^d \ : \ (y-\mu)^\mathsf{T} A (y-\mu) \leq 1\} \ : \quad \mu \in \mathbb{R}^d, \quad A \in \mathbb{R}^{d\times d} \ \ s.p.d. \ \right\}.$$

Then a.s. there exits a $N \in \mathbb{N}$ such that

$$\sum_{i=1}^{n} \mathbf{1}_{\{Y_i \in C\}} \leq \frac{3}{4}\sqrt{n \log\log n} + nP_{Y_1}(C) \quad \text{for all } n \geq N \text{ and all } C \in \mathcal{E}_d \tag{6}$$

Remark 1. *The constant $3/4$ can be replaced by any other constant greater than $\sqrt{2}/2$.*

Proof. Akama and Irie [1] have shown that the VC-dimension of the set $\mathcal{E}_d$ is $(d^2+3d)/2$. From Theorem 1 follows: for any $\varepsilon > 0$ a.s. there exists a $N \in \mathbb{N}$ such that

$$\begin{aligned}
&\sup_{C\in\mathcal{E}_d} \frac{\sum_{i=1}^n 1_C(Y_i) - nP_{Y_1}(C)}{\sqrt{2n\log\log n}} \leq \sup_{c\in\mathcal{E}_d} (P_{Y_1}(C)(1-P_{Y_1}(C)))^{1/2} + \varepsilon \text{ for all } n \geq N \\
&\Rightarrow \\
&\sum_{i=1}^{n} 1_C(Y_i) \leq nP_{Y_1}(C) + (1/2+\varepsilon))\sqrt{2n\log\log n} \text{ for all } n \geq N,\ C \in \mathcal{E}_d.
\end{aligned}$$

□

With the above corollary we can a.s. uniformly bound the number of i.i.d. observations generated by a bounded lebesgue density falling into an elliptical region in $\mathbb{R}^d$.

Corollary 3. *Let $(Y_n)_{n\in\mathbb{N}}$ be i.i.d. variables with a bounded lebesgue density f, $M := \sup_y f(y)$. Then a.s. there exists a $N \in \mathbb{N}$ such that*

$$\sum_{i=1}^{n} \mathbf{1}_{\{(Y_i-\mu)^\mathsf{T}\Sigma^{-1}(Y_i-\mu) \leq (\log|\Sigma|)^2\}} \leq \frac{3}{4}\sqrt{n\log\log n} + \frac{nM\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} |\Sigma|^{\frac{1}{2}} (\log|\Sigma|)^d \tag{7}$$

for every $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$ symmetric positive definite and $n \geq N$.

Proof. First we show $P_{Y_1}(C) \leq M\frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}|\Sigma|^{\frac{1}{2}}(\log|\Sigma|)^d$ for the ellipse $C = \{(y-\mu)^\mathsf{T}\Sigma^{-1}(y-\mu) \leq (\log|\Sigma|)^2\}$ and then we apply Corollary 2.

P_{Y_1} has lebesgue density $f \leq M$. Hence $P_{Y_1}(C) \leq M\lambda^d(C)$. The lebesgue measure of the ellipsoid C is given by $\lambda^d(C) = |\Sigma|^{1/2}\lambda^d(\{y^\mathsf{T}y \leq (\log|\Sigma|)^2\})$ by the invariance of λ^d w.r.t. translations and the substitution rule. For the measure of the sphere it holds $\lambda^d(\{y^\mathsf{T}y \leq (\log|\Sigma|)^2\}) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}(\log|\Sigma|)^d$. □

A bound as in (1) with functions $a(n) = \sqrt{2n\log\log n}$ and $b(n,|\Sigma|) = \frac{nM\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}|\Sigma|^{\frac{1}{2}}(\log|\Sigma|)^d$ is obtained.

4 Conclusion

In the presented work the soft spot in the consistency poof in Chen and Tan [3] was identified, namely the ascription to the univariate case in Lemma 2 there. The introduced alternative in form of Corollary 3 fits almost seamless in Chen's consistency proof. Merely the condition C3 on the penalty function has to be strengthened to $\tilde{p}_n(\Sigma) \leq (\frac{3}{4}\sqrt{n \log\log n}) \log|\Sigma|$ for $|\Sigma| < cn^{-2d}$ for some $c > 0$. However, it is not a problem, since the example penalty function with $\tilde{p}_n(\Sigma) = -n^{-1}(\mathrm{tr}(\Sigma^{-1}) + \log|\Sigma|)$ fulfills this requirement. To see this, assume $|\Sigma| < n^{-2d}$. Then it holds for the eigenvalues of Σ: $\prod_{i=1}^{d} \lambda_i < n^{-2d}$ and hence $\lambda_1 = \lambda_{min} < n^{-2}$. Now, write the trace as the sum of the eigenvalues: $\mathrm{tr}(\Sigma^{-1}) = \lambda_1^{-1} + \ldots + \lambda_d^{-1} > \lambda_{min}^{-1} > n^2$. Finally $-n^{-1}(\mathrm{tr}(\Sigma^{-1}) + \log|\Sigma|) < -n + n^{-1}2d \log n < -(\frac{3}{4}\sqrt{n \log\log n})2d \log n$ for n large enough.

The theoretical background for the new approach is given by Alexander's uniform law of iterated logarithm for VC classes. Elaborate arguments involving Bernstein's inequality and Borel-Cantelli Lemma needed for the one-dimensional case as in Chen et al. [4] are avoided and the proof becomes thereby shorter and more simple.

Moreover, the introduced approach, together with the general proof principle as in Chen and Tan [3] resp. Chen et al. [4] can be used to prove consistency results for penalized MLE for mixtures of distributions with similar properties, like Gamma distributions.

Acknowledgements

I gratefully thank the referees for their helpful and constructive comments and my doctoral supervisor Prof. Dr. Hajo Holzmann for his help in writing the note.

References

[1] Akama, Y. and Irie, K. (2011). VC dimension of ellipsoids. *arXiv:1109.4347 [math.CO]*.

[2] Alexander, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Annals of Probability*, **12** 1041–1067.

[3] Chen, J. and Tan, X. (2009). Inference for multivariate normal mixtures. *Journal of Multivariate Analysis*, **100** 1367–1383.

[4] Chen, J., Tan, X. and Zhang, R. (2008). Inference for normal mixtures in mean and variance. *Statistica Sinica*, **18** 443–465.

[5] Ciuperca, G., Ridolfi, A. and Idier, J. (2003). Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics*, **30** 645–59.

[6] Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Annals of Statistics*, **13** 795–800.

[7] Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, **27** 887–906.

[8] Tanaka, K. (2009). Strong consistency of the maximum likelihood estimator for finite mixtures of location-scale distributions when penalty is imposed on the ratios of the scale parameters. *Scandinavian Journal of Statistics*, **36** 171–184.

[9] van der Vaart, A. and Wellner, J. A. (2000). *Weak Convergence and Empirical Processes*. Springer.

[10] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, **20** 595–601.